
Implicit Media Tagging and Affect Prediction from video of spontaneous facial expressions, recorded with depth camera

By

DANIEL HADAR

Under the supervision of

PROF. DAPHNA WEINSHALL



Department of Cognitive Science
HEBREW UNIVERSITY

A thesis submitted in partial fulfillment of the requirements
for the degree of MASTER OF COGNITIVE SCIENCE in the
Faculty of Humanities.

DECEMBER 2016

ABSTRACT

We present a method that automatically evaluates emotional response from spontaneous facial activity recorded by a depth camera. The automatic evaluation of emotional response, or affect, is a fascinating challenge with many applications, including human-computer interaction, media tagging and human affect prediction. Our approach in addressing this problem is based on the inferred activity of facial muscles over time, as captured by a depth camera recording an individual's facial activity. Our contribution is two-fold: First, we constructed a database of publicly available short video clips, which elicit a strong emotional response in a consistent manner across different individuals. Each video was tagged by its characteristic emotional response along 4 scales: *Valence*, *Arousal*, *Likability* and *Rewatch* (the desire to watch again). The second contribution is a two-step prediction method, based on learning, which was trained and tested using this database of tagged video clips. Our method was able to successfully predict the aforementioned 4 dimensional representation of affect, as well as to identify the period of strongest emotional response in the viewing recordings, in a method that is blind to the video clip being watch, revealing a significantly high agreement between the recordings of independent viewers.

ACKNOWLEDGMENTS

My humble thanks and appreciation to my supervisor, Prof. Daphna Weinshall, that guided me throughout this research and was deeply involved in it. I have been privileged to have her guidance and support.

In addition, I am grateful to Talia Granot, for her practical assistance regarding utilizing facial expressions, as well as our brainstorm meetings that provided me with new ideas and inspiration.

Many thanks also goes to the readers of this dissertation: Hillel Aviezer and Nir Fierstein, for their constructive comments.

Last but not least, I would like to thank my parents, Limor and Shuki, for their never-ending backing and support, and to my beloved wife Liat, for her endless optimism and encouragement.

TABLE OF CONTENTS

	Page
List of Tables	v
List of Figures	vi
1 Introduction	1
2 Theoretical Background and Previous Work	5
2.1 Quantification of Emotion and Facial Expressions	5
2.2 Implicit Media Tagging and Affect Prediction	7
2.3 Depth Cameras	10
2.4 Facial Response Highlight Period	10
3 Database Construction	11
3.1 Eliciting Emotion via Video Clips	11
3.2 Database Criteria	12
3.3 Overview	13
3.4 Method	13
3.5 Results and Analysis	14
4 Method	17
4.1 Experimental Design	17
4.2 Facial Expressions Recording	18
4.3 Features	19
4.4 Predictive Models	21
5 Results and Analysis	25
5.1 Learning Performance	25
5.2 Relative Importance of Features	28
5.3 Localization of Highlight Period	29
6 Discussion	31

TABLE OF CONTENTS

Appendix A Review of Emotion Elicit Databases	33
Bibliography	37

LIST OF TABLES

TABLE	Page
3.1 Emotion elicit database summary.	13
3.2 Mean, median, standard deviation and range of the different scales over all clips, as well as Intra-Correlation Coefficient (<i>ICC</i>) and Cronbach's α	14
4.1 Summary of models learned.	24
5.1 Mean (and std) of Pearson's R between the predicted and actual ranks, as well as the average correlation between the subjective report and media tags.	25
5.2 Accuracy (and std) of the derived binary measure.	27
5.3 Mean (and std) of Pearson's R between the predicted and actual ranks of AP* and One-Viewer-Out method.	27
5.4 Mean error of all viewers subjective ranks and AP-1, IMT-1 and IMT-2's predictions. .	28
A.1 Emotion eliciting video clips databases.	36

LIST OF FIGURES

FIGURE	Page
1.1 The triangular relationship between the facial expression, the media tags and the viewer's affective state.	2
1.2 Facebook's reactions.	2
2.1 Examples from the Facial Action Coding System [28].	6
2.2 Illustration of the difference between affect prediction (f) and implicit media tagging (g).	7
3.1 Correlations between valence, arousal, likability and rewatch.	15
3.2 Distribution of 4 subjective scores over all tested clips, where valence and arousal define the two main axes, also summarized in histogram form above and to the right of the plot. Size and color correspond respectively to the remaining 2 scores of rewatch and likability.	16
4.1 The Experimental design.	18
4.2 (A) Quantized AU signal ($K = 4$), and (B) Its corresponding transition matrix. The number of frames labeled 0,1,2,3 is 6,3,5,5, respectively. Therefore: $ActivationRatio = \frac{13}{19}$, $ActivationLength = \frac{7}{19}$ and $ActivationLevel = 1.473$, and $ChangeRatio = \frac{6}{18}$, $SlowChangeRatio = \frac{4}{18}$, $FastChangeRatio = \frac{2}{18}$	20
4.3 Facial response (middle row) to a video clip (illustrated in the top row), and the time varying intensity of AU12 (bottom row).	21
4.4 Illustration of the two-step prediction algorithm.	22
5.1 Correlation example between predicted and actual ratings of a single viewer's valence score ($R=0.791$).	26
5.2 The relative contribution of different feature groups to IMT-1.	29
5.3 The relative contribution of different feature groups to AP-1.	29
5.4 Histogram of HPs relative to the clips' end time, which marks the origin of the X -axis ($\mu = -7.22$, $\sigma = 4.14$, $\chi^2 = 0.86$).	30

INTRODUCTION

The Roman Philosopher Cicero wrote, "*The face is a picture of the mind*". This statement has been discussed and debated repeatedly over the years, within the scientific community and outside it – are the face really a window to the soul? Does human emotion reflect in facial expressions? There is a vast agreement among scholars that the answer is, at least partially – yes. Hence, we asked the following question – could it be done automatically? That is, could computer vision tools be utilized to evaluate humans' emotional state, given their facial expressions?

In the past two decades we had witnessed an increased interest in automatic methods that extract and analyze human's emotions, or *affective state*. The potential applications of automatic affect recognition vary from human computer interaction to emotional media tagging, including for example the creation of a user's profile in various platforms, building emotion-driven HCI systems, and emotion-based tagging of dating sites, videos on YouTube or posts on Facebook. Indeed, in recent years media tagging has received much attention in the research community (*e.g.* [83, 92]).

In this work we took advantage of the emerging technology of depth cameras. Recently, depth cameras based on structured light technology have emerged as a means to achieve effective human computer interaction in gaming, based on both gestures and facial expressions [1]. We used a depth camera (Carminie 1.09) to record participants facial response to a set of video clips designed to elicit emotional response, and developed two types of pertinent prediction models for automatic quantitative evaluation of affect: models for tagging video clips from human facial expressions (*i.e.* implicit media tagging), and models for predicting viewers affective state, given their facial behavior (*i.e.* affect prediction). Respectively, a clear separation between them should be drawn.

Both implicit media tagging and affect prediction concern the estimation of emotion-related

indicators based on non-verbal cues, but they differ in their target: in the first, the purpose is predicting attributes of *the multimedia stimuli*, while in the latter, *the human affect* is the matter in hand. This distinction could be made clear by observing the triangular relationship between the video clip's affective tagging, the facial response to it and the viewer's reported emotional feedback (see Figure 1.1) – implicit media tagging concerns the automated annotation of a stimuli directly from spontaneous human response, while affect prediction deals with predicting the viewer's affective state. To be noted that objects and locations identification is not a part of this work's scope (*e.g.* [88]), but only emotional-related tags.

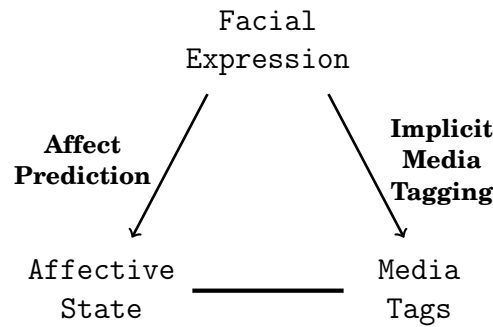


FIGURE 1.1: THE TRIANGULAR RELATIONSHIP BETWEEN THE FACIAL EXPRESSION, THE MEDIA TAGS AND THE VIEWER'S AFFECTIVE STATE.

As opposed to *explicit* tagging, in which the user is actively involved in the tagging process, *implicit* tagging is done passively, and relies only on the typical interaction the user have with such stimuli (*e.g.* watching a video clip). As such, it is less time and energy consuming, and more likely to be free of biases. It has been suggested that explicit tagging tends to be rather inaccurate in practice; for example, users tend to tag videos according to their social needs, which yields tagging that could be reputation-driven, especially in a setup where the user's friends, colleagues or family may be exposed to their tags [77].

In recent years, media tagging had became an integral part of surfing the internet. Many web platforms allow (and even encourage) users to label their content by using keywords (*e.g.* *funny*, *wow*, *LOL*) or designated scales (*e.g.* Facebook's reactions, see Figure 1.2). Clearly, non-invasive methods which produce such tags *implicitly* can be of great interest. That being said, implicit media is complementary (rather than contradictory) to explicit media tagging, and as such can be used for assessing the correctness of explicit tags [45], or for examining the inconsistency between the intentional (explicit) and involuntary (implicit) tagging [63, 64].



FIGURE 1.2: FACEBOOK'S REACTIONS.

Altogether, we learned four different models that share a similar inner-mechanism, but vary in their prior knowledge and target. Specifically, two models were trained to predict the *clip's affective rating* (implicit media tagging), while two other models were trained to predict the *viewer's subjective affective state* for each individual (affect prediction). Specifically, the first model predicts the *affective rating* of a new clip given the facial expressions of a single viewer to a set of known clips, while the second model uses the facial expressions of several viewers. This can be useful for the affective tagging of new video clips by a system relying on the facial expressions of a group of viewers, who had previously been recorded viewing a set of clips with pre-determined affective rating. If no pre-determined affective ratings are available, the third model predicts the *subjective rating* of a new clip, when given the facial expressions and subjective ratings of a single viewer to a set of clips which does not include the new clip.

Affect computation from facial expressions cannot be readily separated from the underlying theories of emotional modeling, which rely on assumptions made by researchers in the field of psychology and sociology, and in some cases are still under debate. Thus the *Facial Action Coding System* (FACS), originally developed by the Swedish anatomist Carl-Herman Hjortsjö [40], is often used to analyze facial activity in a quantitative manner. FACS gives a score to the activity of individual facial muscles called *Action Units* (AUs), based on their intensity level and temporal segments. In our work we used the commercial software *Faceshift* [5] to extract quantitative measures of over 50 AUs from recordings of spontaneous facial activity.

Our goal was to achieve an automatic model that infers descriptive ratings of emotion. To this end we employed the *dimensional approach* to modeling emotion, a framework whose elements are bipolar axes constituting the basis of a vector space (not necessarily orthonormal), and where it is assumed that every emotion can be described as a linear combination of these axes. We represented emotion by a combination of two key scales – *Valence* and *Arousal* (see discussion in Section 2.1). In addition, we added 2 contemporary scales – *Likability* and *Rewatch* (the desire to watch again), which are more suitable for modern uses in HCI and media tagging (following [69]).

Our method is based on learning from a tagged database of video clips. In order to train a successful algorithm and be able to test it against some meaningful ground truth, we needed a database of video clips which can invoke strong emotional responses in a consistent manner across individuals. Since no reliable database of video clips that suites our needs exists at the moment, we constructed a new database from publicly available video clips (see discussion in Chapter 3). This database is available to the community and can be inspected in [2]. A similar empirical procedure was used to collect data for the training and evaluation of our method, as described in Chapter 4.

Next, we developed a vector representation for videos of facial expressions, based on the estimated FACS measurements. Specifically, we started from AUs computed automatically by *Faceshift* [5] from a depth video. The procedure by which we have obtained a concise representation for each video of facial expressions, which involved quantification of both dynamic and spatial features of the AUs activity, is described in Section 4.3.

In Section 4.4 we describe the method by which we learned to predict affect from the ensuing representation for each viewer and each viewed clip. In the first step of our method we generated predictions for small segments of the original facial activity video, employing linear regression to predict the 4 quantitative scales which describe affect (namely *valence*, *arousal*, *likability* and *rewatch*, *a.k.a.* *VALR*). In the second step we generated a single prediction for each facial expressions video, based on the values predicted in the first step for the set of segments encompassed in the active part of the video. The results are described in Chapter 5, showing high correlation between the predicted scores and the actual ones.

There are three main novelties in this work are: first, our models are based directly on the automatically inferred activity of facial muscles over time, considering dozens of muscles, in a method which is blind to the actual video being watched by the subject (*i.e.* the model isn't aware of any of the stimuli details and attributes); Second, the facial muscular activity is measured using only a single depth camera; And third, we present a new publicly available database of emotion eliciting short video clips, which elicit a strong affective response in a consistent manner across different individuals.

THEORETICAL BACKGROUND AND PREVIOUS WORK

Human facial expressions are a fundamental aspect of our every-day lives, as well as a popular research field for many decades. It is not always thought-about or spoken-of, but many of the decisions we take are based upon other people's facial expressions, for example, deciding whether to ask someone's phone number ("is he smiling back?"), or whether a person is to be trusted. Psychologists have been claiming assiduously that human beings are in fact experts of facial expressions (as a car-enthusiast would be an expert of car models), in terms of being able to distinguish between highly similar expressions, recognize large amount of different faces (of friends, family, co-workers, etc.) and being sensitive to subtle facial gestures.

2.1 Quantification of Emotion and Facial Expressions

For over a 100 years now, scholars have been interested in finding a rigorous model for the classification of emotion, seeking to express and explain every emotion by a minimal set of elements. The work in this field can be divided into two approaches: the *categorical approach* and the *dimensional approach* (see [36] for a comprehensive survey). These approaches are not contradictory, as they basically offer alternative ways to describe the same phenomenon. Briefly, the *categorical approach* postulates the existence of several basic and universal emotions (typically surprise, anger, disgust, fear, interest, sadness and happiness), while the *dimensional approach* assumes that each emotion can be described by a point in a 2D/3D coordinate system (typically valence, arousal and dominance).

In this work we collected emotional ratings over the dimensional approach, alongside free language descriptions, for several reasons: first, using *forced choice paradigm* for emotion rating (compelling subjects to chose an emotion from a closed set of basic emotions) may yield spurious and tendentious results, that don't fully grasp the experienced emotion [37]; second, keeping

in mind that mixed feelings can arise from a single stimuli [39, 57], it follows that restricting participants to using a set of discrete emotions might bring about the loss of some emotional depth; and third, basic emotion tags could be extracted from valence and arousal ratings, as well as from the free language text [7, 23].

In the 1970s Ekman and Friesen introduced FACS, that was developed to measure and describe facial movements [27]. The system’s basic components are Action Units (AUs), where most AUs are associated with some facial muscle, and some describe more general gestures, such as sniffing or swallowing (see examples in Fig. 2.1). Every facial expression can be described by the set of AUs that compose it. Over the years, Ekman and his colleagues expanded the set of AUs (and FACS expressiveness accordingly), adding head movements, eye movements and gross behavior. Currently there are several dozen AUs, over 50 of them are solely face-related. Describing a facial expression in terms of AUs is traditionally done by FACS-experts specifically trained for this purpose.

Automated FACS coding that will replace the manual one poses a major challenge to the field of computer vision [74]. AUs extraction can be done using methods based on *geometric features*, such as tracking points or shapes on the face, with features like position, speed and acceleration (e.g., [31, 75, 76]), or using *appearance based* methods based on changes in texture and motion of the skin, including wrinkles and furrows (e.g., [8, 14, 15, 61]). The use of geometric features tends to yield better results, since appearance based methods are more sensitive to illumination conditions and to individual differences, though a combination of both methods may be preferable [96]. A newer promising method is based on temporal information in AU activity, which was found to improve recognition as compared to static methods [53, 65]. Basic features are classified into AUs using model driven methods such as active appearance models (AAMs) [62] or data driven methods [85]. While data driven methods require larger sets of data in order to cope with variations in pose, lightning and textures, they allow for a more accurate and person-independent analysis of AUs [86].

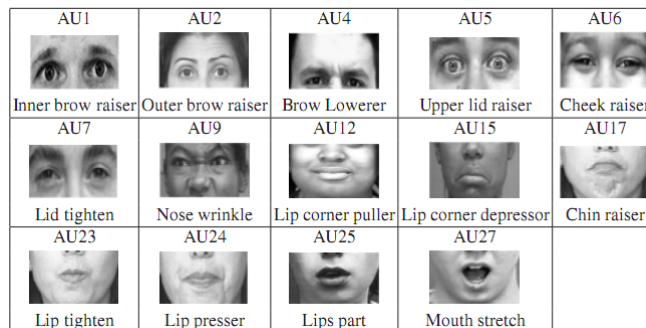


FIGURE 2.1: EXAMPLES FROM THE FACIAL ACTION CODING SYSTEM [28].

2.2 Implicit Media Tagging and Affect Prediction

To our knowledge, tagging of media implicitly and predicting viewers affective state based on data obtained from depth cameras is a recent and uncharted territory. A few exceptions include Niese *et al.* [72] who used evaluated 3D information from the raw data obtained by a 2D camera, and Tron *et al.* [99, 100] who used depth cameras to classify the mental state of schizophrenia patients, based on their facial behavior.

The models in affect prediction are designed to predict each viewer's personal affective state, while in media tagging they attempt to predict a media tag based on input from (possibly different) viewers. Formally, for person p_i who is undergoing the affective state a_i while viewing the clip c_k , an affect prediction model would be defined as: $f(p_i) = a_i$, while an implicit media tagging model would be: $g(p_i) = c_k$, and in particular, for viewers i, j it holds that: $f(p_i) = a_i$, $f(p_j) = a_j$ (when a_i isn't necessarily equals a_j), but $g(p_i) = g(p_j) = c_k$. See illustration in Figure 2.2.

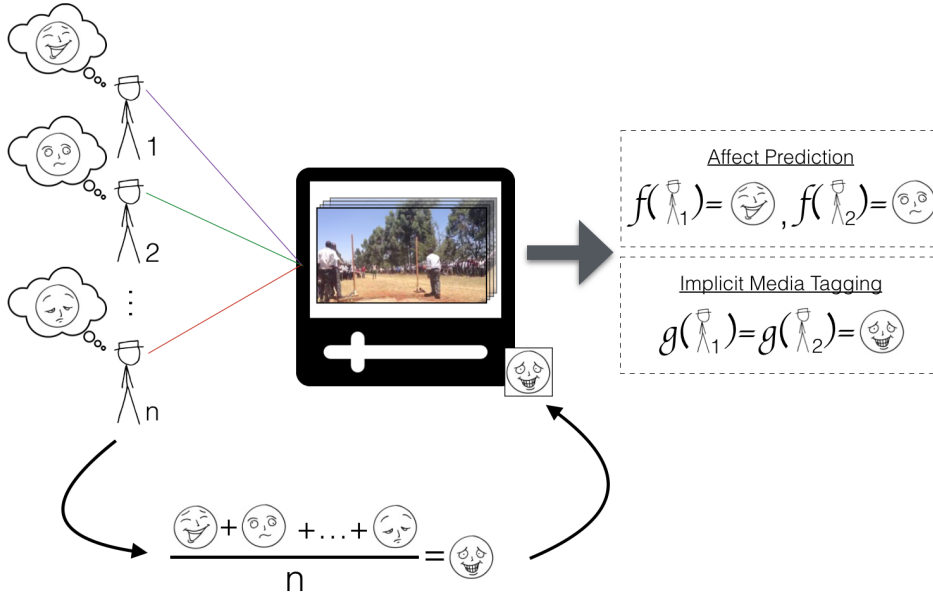


FIGURE 2.2: ILLUSTRATION OF THE DIFFERENCE BETWEEN AFFECT PREDICTION (f) AND IMPLICIT MEDIA TAGGING (g).

2.2.1 Implicit Media Tagging

A number of implicit media tagging theories and models have been developed based on different modalities, including facial expressions, low-level visual features of the face and body, EEG, eye gaze, fMRI, or audio-visual features of the stimuli.

Facial Expressions

Hussein and Elsayed [41] trained a classifier to predict the relevance of a document to a person based on her facial behavior, using 16 feature point along the face. Arapakis *et al.* [9] pursuit a similar goal, but the model's features included the level of basic emotions (e.g., how happy or angry the participant is), combined with other peripheral physiological metrics (such as skin temperature). They achieved higher success rates (accuracy of 66.5% at best) and showed that a user's predicted emotional states can be used to predict the stimuli's relevance.

Jiao and Pantic [45] suggested a facial expressions based model to predict the correctness of tags of images. They used 2D cameras to identify 19 facial markers (such as eyebrows boundaries, lip corners and nostrils) and used Hidden Markov Models to classify their movement along time. Their conclusion was that user's facial reactions convey information about the correctness of tags associated with multimedia data. Similarly, Tkalcic *et al.* [97] presented a method to predict valence, arousal and dominance tags for images, based on the user's facial expressions (given as a sequence of images).

As for video tagging, most works focus on affect prediction. Among the few that discuss media tagging are Zhao *et al.* [113], who proposed a method that based only on the participant's facial expressions to predict categories of movies (e.g. comedy, drama and horror). Almost all papers that use facial expressions for implicit media tagging combine it with additional input, such as content from the video itself or additional physiological measures. Bao *et al.* [12] added the user's acoustic features, motion and interaction with the watching device (tablet or mobile). Wang *et al.* [103] used head motion as well as facial expressions, combined with "common emotions" (the emotions that are likely to be elicited from the majority of users) to predict video's tags, in terms of basic emotions. Another common measurement uses the EEG signal (e.g. [54, 90]).

Other Methods

The use of EEG in this field is relatively common, presumably because it is non-invasive and relatively cheap, permitting the appealing notion that one may be able to tap specific brain localization for different emotional states. EEG is popular for tagging of objects and landscapes (e.g. [30, 49, 51]), but it is also commonly used for emotion-related tagging.

Yazdani *et al.* [110] implemented an EEG-based system that predicts a media's tag, in terms of basic emotions [26]; They achieved an average of 80.19% over a self-developed database of 24 short video clips. Soleymani *et al.* used EEG-based methods to tag clips from MAHNOB-HCI database [91], once combined with pupillary response and eye gaze [94] and once without [93], and reached a success rate of $F1 = 0.56$ for valence and $F1 = 0.64$ for arousal. Wang *et al.* [104] expanded the battery of tools in use, adding features from the video stimuli itself to the EEG recordings, to create a fusion method that achieved an accuracy score of 76.1% over valence and 73.1% over arousal.

Han *et al.* [38] proposed a method that predicts the arousal rating of video clips based on fMRI scans and low-level audio-visual features; this method achieved an average accuracy score of 93.2% for 3 subjects. Jiang *et al.* [43] proposed a framework that interweaves fMRI results to low-level acoustic features and thus enables audio tagging based on fMRI without actually using fMRI further on. Abadi *et al.* [6] combined MEG signal and peripheral physiological signals (such as EOG and ECG) to predict the ratings of movies and music clips, including valence, arousal and dominance ratings, with accuracy score of 0.63, 0.62 and 0.59 (respectively) at best. Similar to [12], in [89] the authors also proposed a method to estimate movie's ratings, based instead on Galvanic Skin Response (GSR).

2.2.2 Affect prediction

Similarly to implicit media tagging, affect prediction models are usually composed of several inputs – some physiological (*e.g.* facial expressions, brain activity or acoustic signals) and some are derived from the media the participant is exposed to (*e.g.* the video clip's tagging or visual and acoustic features). Moreover, they typically rely on the participant's reported emotional state.

Lin *et al.* [59] combined EEG and acoustic characteristics from musical content to evaluate the participant's reported valence and arousal. Similarly, Chen *et al.* [22] used music pieces as a stimuli and combined EEG with subject's gender data to predict valence and arousal. Zhu *et al.* [114] used video clips as stimuli, and also combined between EEG measurements and acoustic/visual features from the clips to predict valence and arousal. Lee *et al.* [58] had used a similar method, but utilized 3D fuzzy GIST for the features extraction and 3D fuzzy tensor for the EEG feature extraction, and predicted valence only. The last two had predicted only a binary result (positive or negative).

As for facial expressions, Soleymani *et al.* [90] compared it to EEG, and found that the results from facial expressions are superior to the results from EEG as a predictor for affective state, as most of the emotionally valuable content in EEG features is a result of facial muscle activity (but EEG signals still carry complementary information). McDuff *et al.* [69] used a narrow set of facial features (namely smiles and their dynamics) to predict participant's likability and desire to watch again of 3 Superbowl ads, with crowdsourced data collected from the participants webcams. Bargal *et al.* [13] were among the first to use deep neural networks on facial expressions (captured by 2D camera) to predict affective states (in terms of basic emotions), and among the few to propose a model based only on facial expressions. In addition, several papers suggest methods to predict participant's level of interest based on their facial expressions, such as Peng *et al.* [80] that implemented a model based on head motion, saccade, eye blinks, and 9 raw facial components; they predicted an emotional score (as well as an attention score) and combined them to predict the level of interest. Arapakis *et al.* [10] proposed a multimodal recommendation system, and showed that the performance of user-profiling was enhanced by facial expression data.

2.3 Depth Cameras

The depth camera used in our study employs IR technology, in which infra-red patterns are projected onto the 3D scene, and depth is computed from the deformations created by 3D surfaces [35]. The technology enables to capture facial surface data, which is less sensitive to head pose and to lightning conditions than 2D data, and yields better recognition of AUs [82, 100]. One drawback, however, is that the depth resolution is rather low, and therefore the image may contain small artifacts in highly reflective and non-reflective regions, or holes in regions not covered by the projector [84].

2.4 Facial Response Highlight Period

One substantial component of our models is identifying the most informative time frame in the participant’s facial response. In fact, during most of the watching time, most participants’ facial behavior showed little to no emotional response; therefore, we sought to find the part of the clip where relevant and informative activity was taking place. Money and Agius [71] distinguished between two types of highlight period localization techniques: internal and external; the first utilized information from the video itself, such as image analysis, objects identification and audio features, while the second analyses information which can be obtained irrespective of the video, such as viewer’s physiological measures or contextual information, the time of the day, the device, or the location in which the viewer is watching the video.

Both internal and external techniques have been exploited to localize a video’s highlight periods. Examples for internal localization are the work of Chan and Jones [21] that extracted affective labels (valence and arousal) from the video’s audio content, and the work of Xu *et al.* [108] that proposed a framework based on the combination of low-level audiovisual features with more progressive features, such as dialogue analysis for informative keywords and emotional intensity.

Our highlight period localization technique is an external one, that is blind to the video clip, and is based solely on the viewers facial expressions (see Section 4.3 for details). Examples of external technique are the works of Joho *et al.* [47, 48] and Chakraborty *et al.* [20]. Joho *et al.* localized highlight periods from viewers facial expressions, based on two feature types: expressions change rate, and pronunciation level – the relative presence of expressions in each of three emotional categories: neutral, low (anger, disgust, fear, sadness) and high (happiness and surprise). Chakraborty *et al.* harnessed highlight periods localization with a model composed of viewers facial expressions and heart rate, in order to detect sports highlights.

DATABASE CONSTRUCTION

This chapter describes the database of emotion eliciting short video clips we have developed. There are currently several available databases of this kind, but unfortunately none of them is suited for our needs. In Section 3.1 we review these databases and the relevant literature, and subsequently we list the criteria that our database must meet (Section 3.2). The full database can be found online [2], also reviewed in Section 3.3. The method we employed to collect ratings for the clips is very similar to the method we used for developing our model, and it is described in Section 3.4 as well as in Chapter 4. The results and their analysis are described in Section 3.5, showing high inter-raters agreement.

3.1 Eliciting Emotion via Video Clips

The traditional stimuli used to elicit emotion in human participants under passive conditions (without active participation by such means as speech or movement) are sound (e.g., [112]), still pictures (e.g., [25, 70]) and video clips (e.g., [16]). Although still pictures are widely used, there are several limitations to this method. First, the stimulus is static, thus it lacks the possibility of examining the dynamics of an evoked emotion [111]. Second, IAPS [56] is by far the most dominant database in use, and therefore almost every emotion-related experiment uses it. Video clips are also not free of issues, but it seems that they are the most suitable for eliciting strong spontaneous authentic emotional response [29, 73, 106]. Thus, for our needs, an emotion eliciting database of video clips is required.

The first pioneers that developed and released affect tagged video clips databases were [81] and [33]. Both used excerpts from well-known films such as *Kramer vs. Kramer* and *Psycho*, and collected ratings from human subjects about their experienced emotion. The considerable

advances in emotional recognition in recent years had motivated scholars to develop and release large batteries of emotion eliciting video clips databases, and we review them in Appendix A.

Initially we defined a set of criteria that our database must meet. However, each of the aforementioned databases lacks at least one of them (as could be seen in Appendix A). We therefore followed the path taken by many other studies that preferred to collect and validate their own databases, or to modify profoundly an existing one (*e.g.* [42, 60, 68, 73, 95, 98, 107]).

3.2 Database Criteria

1. **Duration.** Determining the temporal window's length depends on two main principles: (i) Avoid clips that elicit several distinct emotions in different times; (ii) Have the ability to use many different and diverse clips in a single experiment, without exhausting the subjects. Therefore the clips must be relatively short, but still long enough to elicit a strong clear emotional response.
2. **VALR Rated.** A consequence of our choice of the *dimensional approach* to describe emotion, alongside the scales of *likability* and *rewatch*.
3. **No Sensor Presence.** The awareness of subjects to being observed or recorded (*i.e.* Hawthorne Effect) was found to possibly alter their behavior [24, 66, 67].
4. **Diversity.** Clips should be taken from a variety of domains to reduce the effect of individual variability. Clearly the database should not contain only clips of cute cats in action or incredible soccer tricks, but a balanced mixture.
5. **Globally Germane.** Clips must be intelligible regardless of their soundtrack and content (*e.g.* avoid regional jokes and tales).
6. **Unfamiliarity.** Clips should be such that uninformed viewers are not likely to be familiar with them on one hand, while being publicly available on the other hand.
7. **Not Crowdsourced.** Alongside its strengths, crowdsourcing can be problematic. For example, subjects are less attentive than subjects in a lab with an experimenter alongside them [78], and they differ in their psychological attributes (such as the level of self esteem) from other populations [32]. Moreover, Due to our use of depth cameras, the experiment must be held in a controlled environment (a depth camera is not a household item), and to keep correspondence between the experiment and the database ratings, it should also be tagged in the same environment.
8. **Publicly Available.** To encourage ratification of results and competition, the data must be accessible.

3.3 Overview

The database is composed of 36 short publicly available video clips (6-30 seconds, $\mu = 20_{sec.}$). Each clip was rated by 26 participants on 5 scales, including valence, arousal, likability, rewatch (the desire to watch again) and familiarity, and in addition it was verbally described. Table 3.1 gives a summary of the database.

Number of Clips	36
Duration (in seconds)	6-30 ($\mu = 20.0$)
Number of Raters	26 (13 males and 13 females)
Available Scales	Valence [1-5] Arousal [1-5] Likability [1-3] Rewatch [1-3] Familiarity [0, 1-2, 3+] Free Text [Hebrew]

TABLE 3.1: EMOTION ELICIT DATABASE SUMMARY.

3.4 Method

A highly similar empirical framework was applied in both phases of data collection in this work, database construction and models development. Specifically, the depth camera was only present in the second phase. Here we only describe in details the clips selection process, and the methodology we adopted for the assessment phase is described in Chapter 4.

Clips Selection Over a 100 clips were initially selected from online video sources (such as YouTube, Vimeo and Flickr), to be eventually reduced to 36. We attempted to achieve a diverse set of unfamiliar clips, and therefore focused on lightly viewed ones. We excluded clips that might offend participants by way of pornography or brutal violence¹. Several clips were manually curtailed to remove irrelevant content, scaled to fit a 1440 x 900 resolution, and balanced to achieve identical sound volume.

Clips Assessment 26 volunteers with normal vision participated in the study, for which they received a small payment (13 males and 13 females between the ages 19-29, $\mu = 23.5$). For the method employed, see Chapter 4.

¹As a rule of thumb, we used videos that comply with the YouTube's Community Guidelines [4].

3.5 Results and Analysis

All clips were found to be significantly unfamiliar across all raters ($p < .0001$), and no influence of gender was found. Moreover, ratings on all scales showed high inter-rater agreement with an average Intra-Correlation Coefficient (ICC) of 0.945 (two-way mixed model, $CI = .95$). The results are illustrated in Figure 3.2, and detailed in Table 3.2.

	Mean	Median	STD	<i>min</i>	<i>max</i>	ICC	α
Valence	3.04	3.21	1.00	1.23	4.42	0.975	0.973
Arousal	3.08	3.02	0.65	1.73	4.19	0.926	0.923
Likability	2.02	2.04	0.56	1.12	2.92	0.954	0.952
Rewatch	1.73	1.75	0.44	1.08	2.54	0.927	0.924

TABLE 3.2: MEAN, MEDIAN, STANDARD DEVIATION AND RANGE OF THE DIFFERENT SCALES OVER ALL CLIPS, AS WELL AS INTRA-CORRELATION COEFFICIENT (*ICC*) AND CRONBACH’S α .

There were strong correlations between several scales, most notably valence–likability (Pearson’s $R = .92$), valence–rewatch ($R = .87$) and likability–rewatch ($R = .94$). Interestingly, no significant correlation was found between arousal and likability ($R = -.23$) or between arousal and rewatch ($R = -.04$); a possible explanation could be that some high arousal clips could be very pleasing (such as hilarious clips), while others are difficult to watch (like car accident commercials), as opposed to high valence clips that are unlikely to discontent anyone. As for valence–arousal, a small negative correlation was found ($R = -.40$), possibly because clips with extremely high V-A values that mostly included pornographic content were excluded, although this result does correspond to prior findings [34]. The results are shown in Figure 3.1.

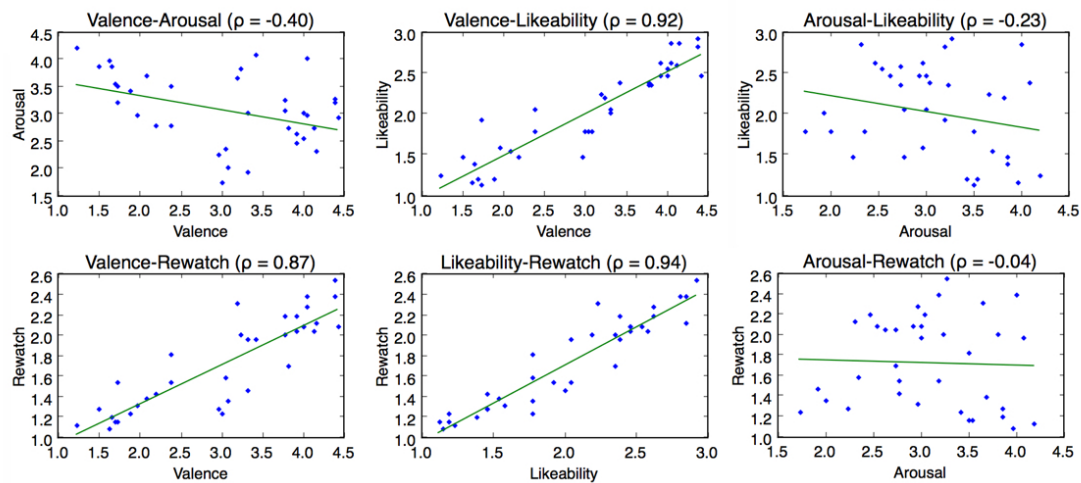


FIGURE 3.1: CORRELATIONS BETWEEN VALENCE, AROUSAL, LIKABILITY AND REWATCH.

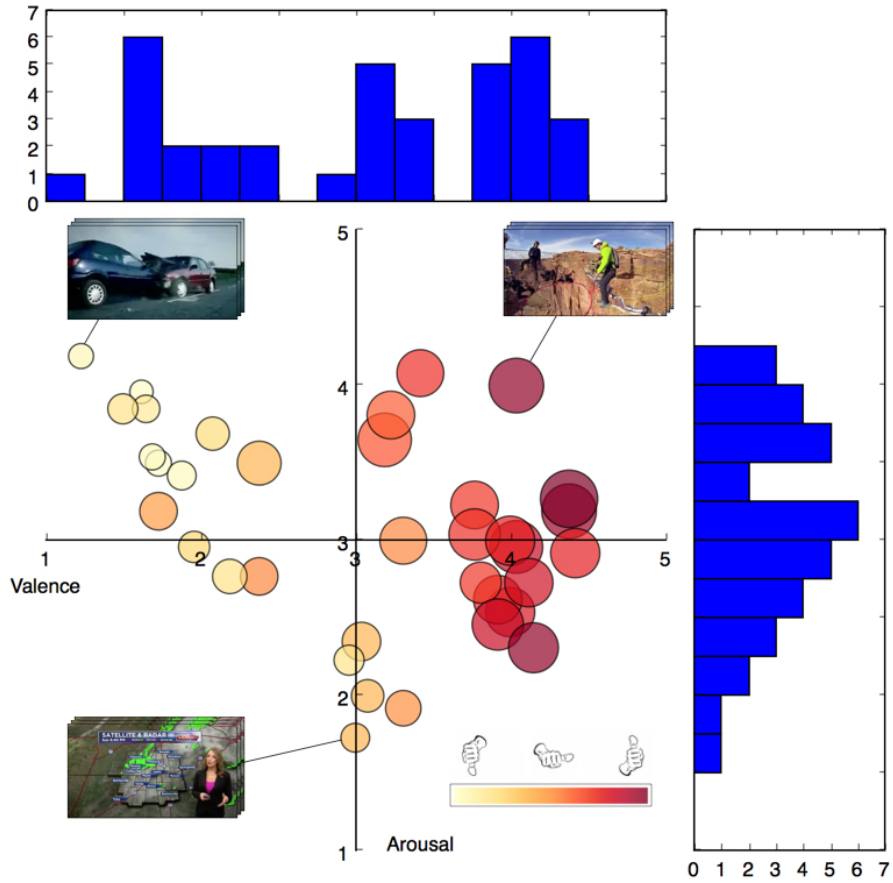


FIGURE 3.2: DISTRIBUTION OF 4 SUBJECTIVE SCORES OVER ALL TESTED CLIPS, WHERE VALENCE AND AROUSAL DEFINE THE TWO MAIN AXES, ALSO SUMMARIZED IN HISTOGRAM FORM ABOVE AND TO THE RIGHT OF THE PLOT. SIZE AND COLOR CORRESPOND RESPECTIVELY TO THE REMAINING 2 SCORES OF REWATCH AND LIKABILITY.

Data collection had two phases: (i) Collect and evaluate a suitable database of video clips which elicit strong and consistent emotional responses in their viewers, as described in Chapter 3; (ii) Record peoples spontaneous facial expressions when viewing these clips. As mentioned, a highly similar empirical framework was used in both phases. In Section 4.1 we describe the experimental environment and workflow. In Section 4.2 we describe the data collection process of the raw recordings, which was used later to calculate the features used in our models, and is elaborated in Section 4.3. Following the assemblage of features we learned our prediction models, a process that we describe in Section 4.4.

4.1 Experimental Design

Participants Data collection was carried out in a single room, well lit with natural light, with minimal settings (2 tables, 2 chairs, a whiteboard, 2 computers and no pictures hanging on the walls). Participants were university students recruited using banners and posters. 26 volunteers with normal vision (13 males and 13 females between the ages 19-29, $\mu = 23.5$ in phase 1, 14 males and 12 females between the ages 20-28, $\mu = 23.3$ in phase 2) participated in this study, for which they received a small payment.

Data Collection Each data collection session consisted of the following stages (see Figure 4.1):

1. A fixation cross was presented for 5 seconds, and the participant was asked to stare at it.
2. A video clip was presented.

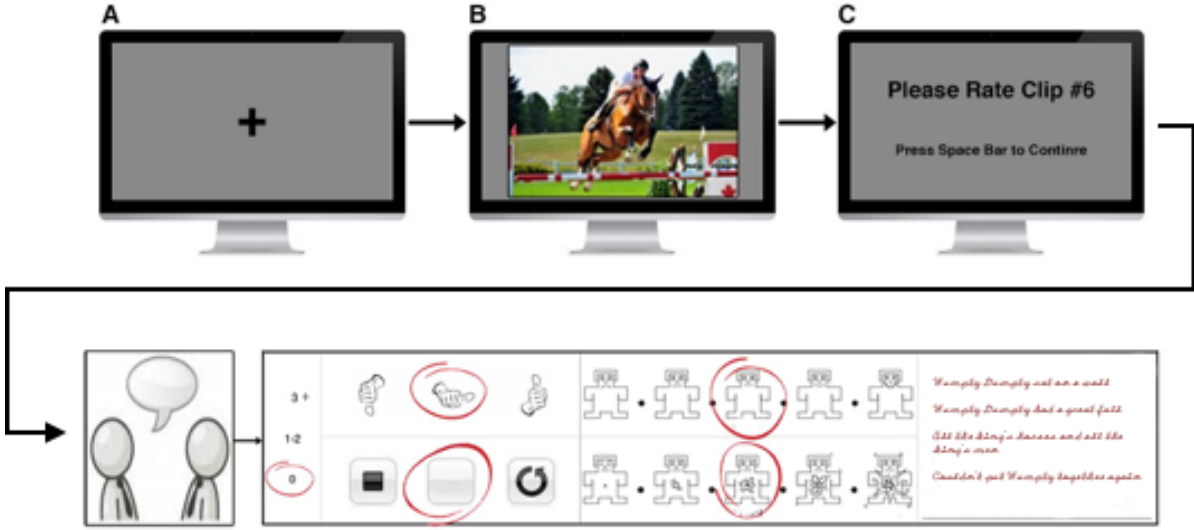


FIGURE 4.1: THE EXPERIMENTAL DESIGN.

3. The participant described verbally his subjective emotion to the experimenter, using two sentences at most.
4. The participant rated her subjective feelings on a pre-printed ratings paper (following [19]).

Data collection was carried out via a self-written Matlab program, designed using Psychophysics Toolbox Extension [18, 50, 79]. Five discrete scales were used for rating: *valence*, *arousal*, *likability*, *rewatch* (the desire to watch again) and *familiarity*, alongside free text description in Hebrew. Specifically, we used SAM Manikins [55] for valence and arousal, the "liking scale" for likability [52], and self-generated scales for rewatch and familiarity (see Figure 4.1). The SAM Manikins method was chosen because it is known to be parsimonious, inexpensive and quick, as well as comprehensive [17]. After every 4 trials, a visual search task was presented ("Where's Waldo?") in order to keep the participants focused, and to force them to change their head situs, sitting position and focal length. The clips order was randomized, and the entire procedure lasted for about an hour. We encouraged participants to rate the clips according to their perceived, inner and subjective emotion. In addition, each participant completed the *Big-Five Personality Traits* questionnaire [46].

4.2 Facial Expressions Recording

18 of the 36 clips were selected from the database. Aiming for a diverse corpus, we chose clips whose elicited response spanned the spectrum of *VALR* as uniformly as possible, also favoring clips with high intra-rater agreement.

Each participant’s facial activity was recorded during the entire procedure, using a 3D structured light camera (Carminc 1.09). Participants were informed of being recorded and signed a consent form. 21 of the 26 participants reported after the experiment their belief that they were not affected by the recording, and that in fact they had forgotten of being recorded. Moreover, the subjective reports in the second phase on all 4 scales had a very similar distribution to the reports in the first phase: Valence ($R = .98$, $p < .0001$), Arousal ($R = .90$, $p < .0001$), Likability ($R = .97$, $p < .0001$) and Rewatch ($R = .95$, $p < .0001$). We therefore believe that the video affect tagging obtained in the database assemblage phase remains a reliable predictor of the emotional response elicited in the recorded experiment as well.

4.3 Features

Previous work in this field calculated facial features either by extracting raw movement of the face without relating to specific facial muscles (e.g., [103]), or by extracting the activity level of a single or a few muscles (e.g., [69, 101, 109]). In this work we extracted the *intensity signals* of over 60 AUs and facial gestures; this set was further analyzed manually to evaluate tracking accuracy and noise levels. Eventually 51 of this set of AUs were selected to represent each frame in the clip for further analysis and learning, including eyes, brows, lips, jaw and chin movements (see example in Fig. 4.3).

Using this *intensity level* representation, providing a time series of vectors in \mathbb{R}^{51} , we computed higher order features representing the facial expression more concisely. This set of features can be divided into 4 types: *Moments*, *Discrete States*, *Dynamic* and *Miscellaneous*.

Moments. The first 4 moments (mean, variance, skewness and kurtosis) were calculated for each AU in each facial video recording.

Discrete States Features. For each AU separately, the raw intensity signal was quantized over time using *K-Means* ($K = 4$), and the following four facial activity characteristic features were computed (see Figure 4.2 for an example):

- **Activation Ratio:** Proportion of frames with any AU activation.
- **Activation Length:** Mean number of frames for which there was continuous AU activation.
- **Activation Level:** Mean intensity of AU activation.
- **Activation Average Volume:** Mean activation level of all AUs, was computed once for each expression.

Dynamic Features. A transition matrix M was generated, measuring the number of transitions between the different levels described above, and three features were calculated for each AU based on it (see Figure 4.2 for an example):

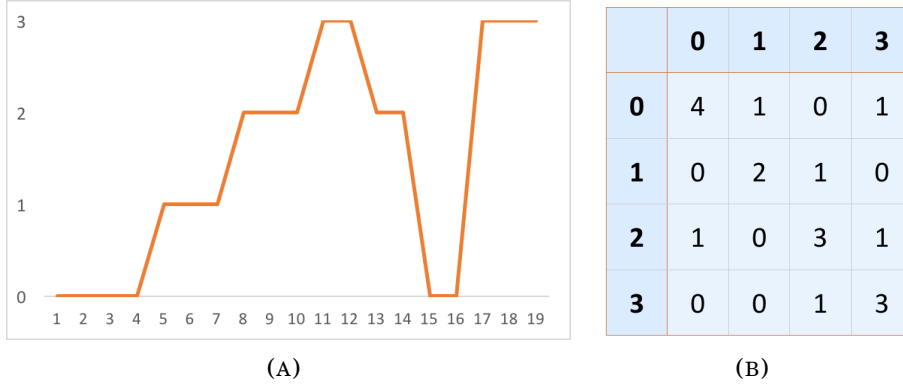


FIGURE 4.2: (A) QUANTIZED AU SIGNAL ($K = 4$), AND (B) ITS CORRESPONDING TRANSITION MATRIX. THE NUMBER OF FRAMES LABELED 0,1,2,3 IS 6,3,5,5, RESPECTIVELY. THEREFORE: $ActivationRatio = \frac{13}{19}$, $ActivationLength = \frac{7}{19}$ AND $ActivationLevel = 1.473$, AND $ChangeRatio = \frac{6}{18}$, $SlowChangeRatio = \frac{4}{18}$, $FastChangeRatio = \frac{2}{18}$.

- **Change Ratio:** Proportion of transitions with level change.
- **Slow Change Ratio:** Proportion of small changes (difference of 1 quantum).
- **Fast Change Ratio:** Proportion of large changes (difference of 2 quanta or more).

Miscellaneous Features. Including the number of smiles and blinks in each facial response. The amount of smiles was calculated by taking the maximum of the amount of peaks in the signals of both lip corners, where *peak* is defined as a local minimum which is higher by at least 0.75 as compared to its surrounding points. The amount of blinks was calculated in a similar manner, with a threshold of 0.2.

Highlight Period

In most video clips, during most of the viewing time, participants' facial activity showed almost no emotional response. Respectively, considering the entire duration of the facial expression when calculating features is not only unnecessary, but could actually harm the model as it adds noise to it. We therefore sought to find the time frame of each facial expression in which the relevant activity (in terms of affective response) was taking place. Specifically, we implemented a model that localized the *highlight period* solely from the viewer's facial expression, in a technique that is blind to the video clip.

For each participant and clip, our model receives his/hers muscular intensity levels for the clip's duration (with 6 seconds margins from its beginning and end), and isolates the activity of gestures we found to be most informative (namely smiles, blinks, mouth dimples, lips stretches and mouth frowns). Later it localizes the 6 seconds window in which these gestures achieves maximal average intensity and variance. Excluding moments, All features were computed based

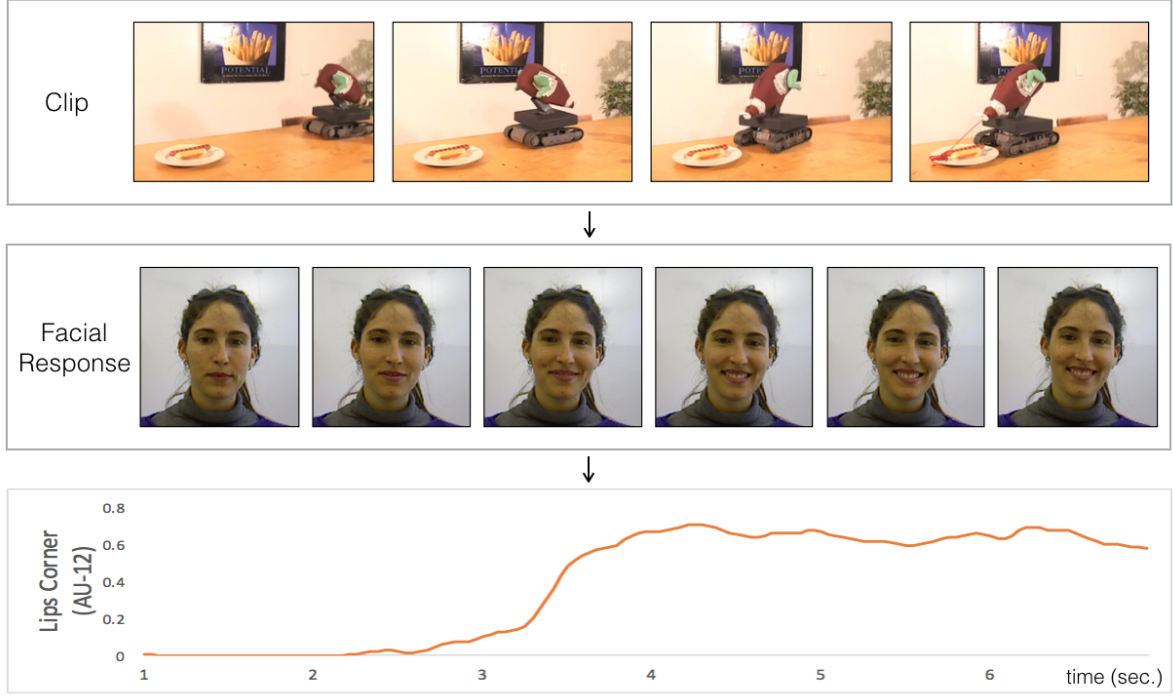


FIGURE 4.3: FACIAL RESPONSE (MIDDLE ROW) TO A VIDEO CLIP (ILLUSTRATED IN THE TOP ROW), AND THE TIME VARYING INTENSITY OF AU12 (BOTTOM ROW).

only on the highlight period. Notice that for some clip C_i , the highlight period might be different for every pair of subjects (although its duration will be the same, as it is an simplifying assumption of our model).

4.4 Predictive Models

In this final step we learned two types of prediction models – implicit media tagging models (*IMT*), and affect prediction models (*AP*). Both model types predict an affective rank (in VALR terms) when given as input a facial expression recording, represented by a vector in \mathbb{R}^d feature space ($d = 462$). Since the number of participants in our study was only 26, a clear case of small sample, the full vector representation – if used for model learning – would inevitably lead to overfit and poor prediction power. We therefore started by significantly reducing the dimension of the initial representation of each facial activity using PCA. Our final method employed a two-step prediction algorithm (see illustration in Figure 4.4), as follows:

First step After the highlight period of each clip was detected (for each subject), it was divided into n fixed size overlapping segments. A feature vector was calculated for each segment, and a linear regression model (f_1) was trained to predict the 4-dimensional affective scores

vector of each segment.

Second step Two indicators (mean and std) of the set of predictions over all segments in the clip were calculated, and another linear regression model (f_2) was trained to predict the 4-dimensional affective scores from these indicators.

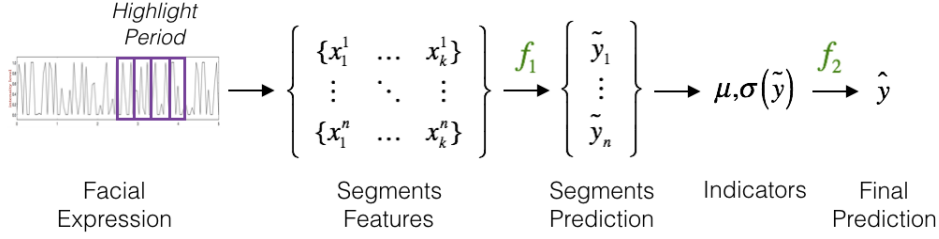


FIGURE 4.4: ILLUSTRATION OF THE TWO-STEP PREDICTION ALGORITHM.

Several parameters control the final representation of each facial expression clip, including the number of segments, the length of each segment, the percent of overlap between the segments, the final PCA dimension and whether PCA was done over each feature type or over all features combined. The values of these parameters were calibrated using cross-validation: given a training set of l points, the training and prediction process was repeated l times, each time using $l - 1$ points to train the model and predict the value of the left out point. The set of parameters which achieved the best average results over these l cross validation repetitions was used to construct the final facial expression representation of all datapoints.

Different types of predictive models

Altogether, we learned four different models that shared this mechanism, but varied in their prior knowledge and target (see Table 4.1 for formal definition). Specifically, the first two models were trained to predict the *clip's affective rating* as stored in the database (implicit media tagging), while the last two models were trained to predict the *viewer's subjective affective state* for each individual (affect prediction).

Implicit media tagging of unseen clips (IMT-1). This model is built using the facial expressions of a single viewer. Given the facial response of this viewer to a **new clip**, the model predicts the *clip's affective rating*.

Implicit media tagging of unseen clips via multiple viewers (IMT-2). Given the facial response of a *set of familiar viewers* to a **new clip**, the model predicts the *clip's affective rating*. Generalizing the first model, this second model predicts the new clip's affective rating by taking into account the prediction of *all* viewers.

Viewer’s affect prediction for an unseen clip (AP-1). This model is built using the facial expressions of a single viewer. Given the facial response of this viewer to a **new clip**, the model predicts the *viewer’s subjective affective state*.

Affect prediction of new viewers (AP*). This model is built separately for each clip, using the facial expressions of all the viewers who had watched this clip. Given the facial response of a **new viewer**, the model predicts the new viewer’s *subjective affective state* when viewing this clip.

Formally, denote the following:

- Let e_i^j denote the facial expression representation of viewer i to clip j .
- Let v_i^j (*respectively*: a_i^j, l_i^j, r_i^j) denote the valence (*respectively*: arousal, likability, rewatch) score of viewer i to clip j .
- Let $s_i^j \equiv (v_i^j, a_i^j, l_i^j, r_i^j)$ denote the affective vector score of viewer i to clip j .
- Let \mathbf{v}^j (*respectively*: $\mathbf{a}^j, \mathbf{l}^j, \mathbf{r}^j$) denote the valence (*respectively*: arousal, likability, rewatch) rating of clip j .
- Let $\mathbf{c}^j \equiv (\mathbf{v}^j, \mathbf{a}^j, \mathbf{l}^j, \mathbf{r}^j)$ denote the affective vector rating of clip j .

Model Name	Input	Model's supervision	Output	Comment
IMT-1	e_m^k	$\left\{e_m^j\right\}_{j=1, j \neq k}^{18}$ $\mathbf{c}^1 \dots \widehat{\mathbf{c}^k} \dots \mathbf{c}^{18}$	\mathbf{c}^k	A unique model is built for each viewer m . Note that the model is utterly unfamiliar with clip k .
IMT-2	$\{e_i^k\}_{i \in S, S \subset [26]}$	$\left\{\left\{e_i^j\right\}_{j=1, j \neq k}\right\}_{i=1}^{26}$ $\mathbf{c}^1 \dots \widehat{\mathbf{c}^k} \dots \mathbf{c}^{18}$	\mathbf{c}^k	The output is a single model, which predicts the clip's rank (\mathbf{c}^k) based on the average of the m models computed in IMT-1.
AP-1	e_m^k	$\left\{\left(e_m^j, s_m^j\right)\right\}_{j=1, j \neq k}^{18}$	s_m^k	Similarly to IMT-1, a unique model is built for each viewer m . The model is also unfamiliar with clip k .
AP*	$\left\{e_m^j\right\}_{j=1}^{18}$	$\left\{\left\{\left(e_i^j, s_i^j\right)\right\}_{j=1}^{18}\right\}_{i=1, i \neq m}^{26}$	$\left\{s_m^j\right\}_{j=1}^{18}$	The model predicts s_m^j separately for each clip j , and is also utterly unfamiliar with viewer m .

TABLE 4.1: SUMMARY OF MODELS LEARNED.

RESULTS AND ANALYSIS

To evaluate the predictive power of our models, we divided the set of recordings following a Leave-One-Out (LOO) procedure. Specifically, for IMT-1, IMT-2 and AP-1 we trained each model based on $n - 1$ clips ($n = 18$) and tested our prediction on the clip which was left out. For AP* the procedure was identical, up to leaving a viewer out (instead of a clip), hence $n = 26$. The results are shown in Section 5.1, followed by an analysis of the relative importance of the different feature types (moments, discrete state features, dynamic features and miscellaneous features) in Section 5.2, and the localization of highlight period in Section 5.3.

5.1 Learning Performance

Learning performance was evaluated by *Pearson's R* between the actual VALR scores and the models' predicted ones (see example in Figure 5.1). Table 5.1 shows the average VALR results over all clips/viewers (all correlations are significant, $p < 0.0001$).

	Valence	Arousal	Likability	Rewatch
IMT-1	.752 (.14)	.728 (.07)	.637 (.22)	.661 (.15)
IMT-2	.948 (.22)	.874 (.22)	.951 (.17)	.953 (.19)
AP-1	.661 (.17)	.638 (.19)	.380 (.16)	.574 (.19)
AP*	.561 (.26)	.138 (.21)	.275 (.23)	.410 (.17)
Report/Tags	.783 (.08)	.461 (.33)	.659 (.13)	.561 (.23)

TABLE 5.1: MEAN (AND STD) OF PEARSON'S R BETWEEN THE PREDICTED AND ACTUAL RANKS, AS WELL AS THE AVERAGE CORRELATION BETWEEN THE SUBJECTIVE REPORT AND MEDIA TAGS.

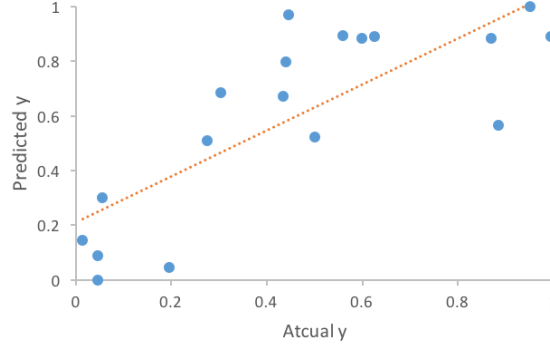


FIGURE 5.1: CORRELATION EXAMPLE BETWEEN PREDICTED AND ACTUAL RATINGS OF A SINGLE VIEWER’S VALENCE SCORE ($R=0.791$).

Notably, implicit media tagging models reached higher success rates than the affect prediction ones. In particular, although IMT-1 and AP-1 both predict an affective state when given a familiar viewer’s facial response to a new clip, the first yields noticeably higher success rates ($\mu = .131$). Based on these results, it could be suggested is that it is easier to predict the *expected* affective state from viewer’s facial expressions than the *reported* one. This is rather surprising, as the opposite sounds more likely – that one’s facial behavior would be a better predictor to his/hers *subjective* emotion than their *expected* one. Under the assumption that the participants in this research proclaimed the true experienced emotion they underwent (as requested), it implies that human facial expressions are more faithful to the actual inner emotion than to the one reported. One reservation is that the viewer’s subjective report is given on a discrete scale, while the media is tagged on a continuous one (as these are average ratings), thus the correlation is likely to be higher for the latter. This problem is averted by comparing the results after binarizing the predictions.

Binary prediction

Apart from the aforementioned reason, often what is needed in real-life applications is a discrete binary prediction rather than a continuous grade, in order to indicate, for example, whether the viewer likes a video clip – or not. To generate such a measure, we binarized the actual ratings and the predicted ones, using the corresponding mean of each measure as a threshold, and calculated the accuracy score (ACC) between them. Since the scores around the mean are rather ambiguous, we eliminated from further analysis the clips whose original tag was uncertain. This included clips with scores in the range $\mu \pm \sigma$, where μ denotes the average score over all clips and σ its *std*, thus eliminating 15% on average of all data points. The results can be found in Table 5.2.

Comparing the results of IMT-2 to published state-of-the-art methods (*e.g.* [69]) demonstrates great success, as an ability to tag media in VALR terms with accuracy rates ranging around 90% is unprecedented. That being said, it is important to note that these results were obtained using

	Valence	Arousal	Likability	Rewatch
IMT-1	71% (.15)	56% (.14)	68% (.12)	73% (.13)
IMT-2	94% (.20)	90% (.18)	91% (.21)	91% (.22)
AP-1	70% (.17)	53% (.22)	49% (.18)	49% (.27)
AP*	64% (.22)	50% (.17)	40% (.25)	42% (.26)

TABLE 5.2: ACCURACY (AND STD) OF THE DERIVED BINARY MEASURE.

a newly composed database (presented in Chapter 3), therefore further research must be carried out for balanced comparison between methods.

Regardless to whether the scale is discrete, the ratio between the analogous algorithms (namely IMT-1 and AP-1) remains similar. Hence, these results support the aforementioned claim that human facial behavior is more faithful to the actual inner emotion than to the one reported. Furthermore, facial expressions allow for better predictions of media tags than the viewer’s subjective rating, as the success rates of IMT-1 are generally higher than the correlation between the viewer’s reported affective state and the media tags (see *Report / Tags* is Table 5.1). These findings are supported by theories of emotional self-report bias, that could arise from many factors (such as cultural stereotypes, the presence of an experimenter, and the notion of the reports being made public).

Unfortunately, the binary predictions of AP* are generally no better than random predictions. Furthermore, for continuous ranks, the algorithm’s predictive power is limited; when a model supervises a group of viewers (like AP*), it is preferable to predict a new viewer’s affective state using only their *affective reports*, without even relying on their facial behavior, as the *average* affective rank of $n - 1$ viewers provides a more accurate prediction to the n -th viewers (namely One-Viewer-Out method), as could be seen in Table 5.3. In other words, when given a group of viewers’ facial expressions and affective ranks, using the average of their ranks would yield better prediction for a new viewer’s rank than his/hers facial expression.

	Valence	Arousal	Likability	Rewatch
AP*	.561 (.26)	.138 (.21)	.275 (.23)	.410 (.17)
Average ($n - 1$)	.779 (.08)	.472 (.29)	.645 (.13)	.552 (.21)

TABLE 5.3: MEAN (AND STD) OF PEARSON’S R BETWEEN THE PREDICTED AND ACTUAL RANKS OF AP* AND ONE-VIEWER-OUT METHOD.

In addition, although seemingly AP-1 yields rather accurate predictions, deeper inspection reveals that an alternative mechanisms based on IMT-1 could be suggested, namely return the predictions of IMT-1 instead; this mechanism yields more accurate predictions of the *subjective affect rank* than AP-1. In other words, if the media’s tags are also available, it’s preferable to train

a model to predict these tags and rely on this model alone, rather than on a model trained to predict the viewer’s subjective rank. Moreover, if other viewers data is also available, relying on it (namely returning IMT-2’s predictions) is even more superior. We demonstrate this observation by inspecting the mean error (ME) between the actual subjective ranks and the predictions given by the AP-1, IMT-1 and IMT-2 models, see Table 5.4.

	Valence	Arousal	Likability	Rewatch
AP-1	.265	.370	.388	.371
IMT-1	.214	.262	.325	.341
IMT-2	.179	.239	.295	.306

TABLE 5.4: MEAN ERROR OF ALL VIEWERS SUBJECTIVE RANKS AND AP-1, IMT-1 AND IMT-2’S PREDICTIONS.

5.2 Relative Importance of Features

We analyzed the relative importance of the different facial features for IMT-1, observing that different facial features contributed more or less, depending on the affective scale being predicted. Features’ relative importance was calculated by learning the models as described above, while using only a single type of features at each time, and comparing the predictions’ success rates. For example, we observed that for IMT-1 the prediction of *valence* relied on all 4 feature types (including moments, discrete state features, dynamic features and miscellaneous features), while the prediction of *arousal* didn’t use the miscellaneous features at all, but relied heavily on the dynamic aspects of the facial expression. Similarly, the prediction of *likability* utilized the miscellaneous features the most, while not using the moments features. Specifically, prediction with only miscellaneous features achieved correlation of $R = 0.275$ with the *likability* score, and $R = 0.287$ with the *rewatch* score. These observations are summarized in Figure 5.2.

As a comparison, we analyzed relative importance of the different facial features for the analogous affect prediction algorithm (namely AP-1). As can be seen in Figure 5.3, the distribution is similar in both models, except that the dynamic features’ relative importance is consistently higher in affect prediction ($\mu = +11.75\%$). Because the dynamic features are relatively more dominant in predicting subjective emotion, this observation encourages us to hypothesize that the temporal aspects of viewers’ facial behavior, in addition to serving as a good predictor for emotions in general, could be used as a distinguishing property between different viewers. This idea is in line with [100], where it was shown that these aspects helps distinguishing between schizophrenia patients and healthy individuals.

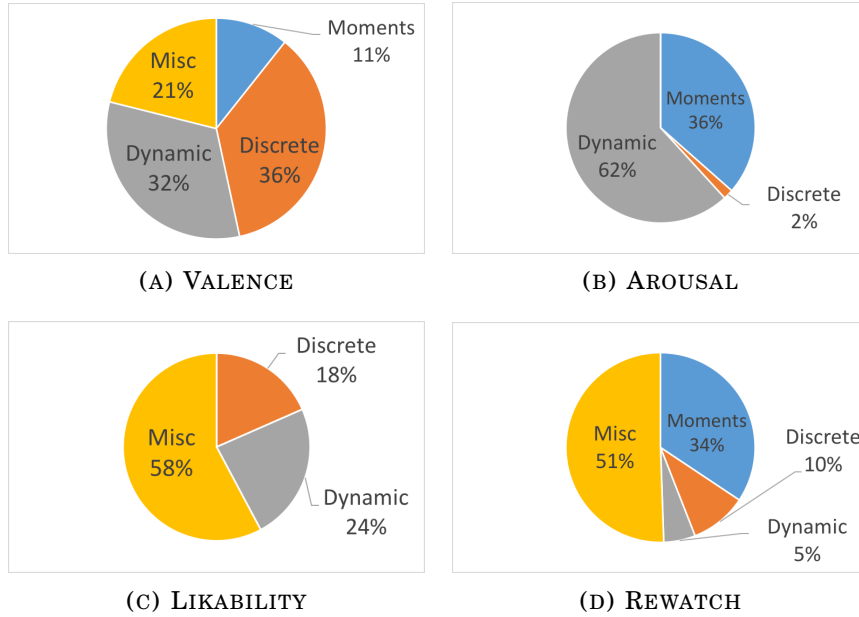


FIGURE 5.2: THE RELATIVE CONTRIBUTION OF DIFFERENT FEATURE GROUPS TO IMT-1.

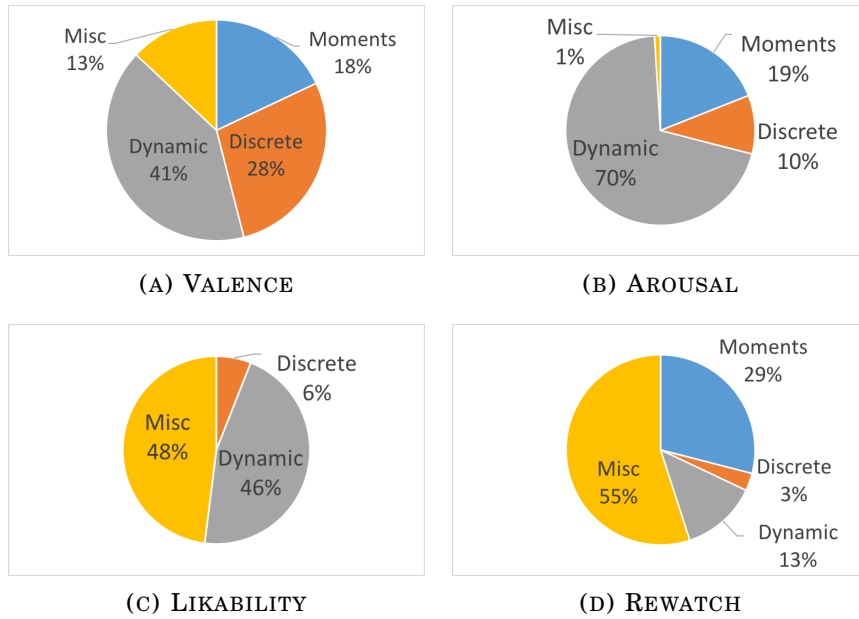


FIGURE 5.3: THE RELATIVE CONTRIBUTION OF DIFFERENT FEATURE GROUPS TO AP-1.

5.3 Localization of Highlight Period

We also analyzed the relative location of the response highlight period (HP) within the clip. Although this period was computed bottom-up from the facial recording of each individual viewer and without access to the observed video clip, the correspondence between subjects was

notably high ($ICC = .941$). Not surprisingly, the beginning of the period was usually found a few seconds before the clip's end ($\mu = -7.22$, $\sigma = 4.14$), and in some clips it lasted after the clip ended (specifically in 8 out of the 18 clips). Yet the HP localization clearly depended, in a reliable manner across viewers, on the viewed clip. For example, when viewing a car safety clip, the average HP started 14 seconds before its end, probably because a highly unpleasant violent sequence of car crashes had begun a second earlier. We may conclude that the HP tends to focus around the clip's end most of the times, but clip-specific analysis is preferable in order to locate it more precisely. The distribution of HPs is presented in Figure 5.4.

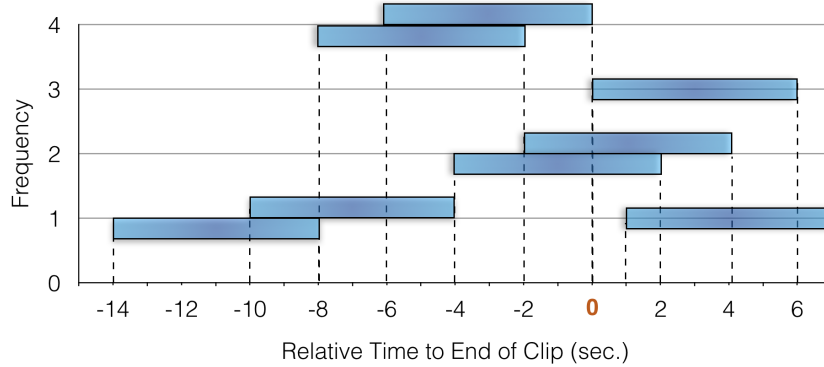


FIGURE 5.4: HISTOGRAM OF HPs RELATIVE TO THE CLIPS' END TIME, WHICH MARKS THE ORIGIN OF THE X -AXIS ($\mu = -7.22$, $\sigma = 4.14$, $\chi^2 = 0.86$).

DISCUSSION

Our contribution in this work is two fold. First, we obtained a database of video clips which give rise to strong predictable emotional response, as verified by an empirical study, and which is available to the community. Second and more importantly, we described several algorithms that can predict emotional response based on spontaneous facial expressions, recorded by a depth camera. Our method provided a fairly accurate prediction for 4 scores of affective state: *Valence*, *Arousal*, *Likability*, and *Rewatch* (the desire to watch again). We achieved high correlation between the predicted scores and the affective tags assigned to the video. In addition, our results suggest that a group of viewers performs better as a predictor to media tagging than a single viewer (similarly to idea of "The Wisdom of Crowds"), as IMT-2 achieves evident higher success rates than IMT-1. Hence, in real-life systems, it's preferable to rely on the facial behavior of a group of known viewers than a single one, if possible.

When using facial expressions for automatic affect prediction, We saw that it's easier to predict the *expected* affective state (*i.e.* implicit media tagging) than the viewer's *reported* affective state (*i.e.* affect prediction). Further analysis evinced that a prediction based on the media tags provides a more accurate estimation for the viewers' affective state than a prediction based on the viewer's report. These results are rather surprising, and we believe that further effort to improve the affect prediction models could obtain better results. One possible course of action to expand the predictive power is utilizing viewers personality properties by using the Big-5 Personality Traits [46], that were also collected in this study. Another approach is considering head and upper-body gestures, that could be also captured by depth cameras, as it has been showed that body movements contribute to emotion prediction [11].

Interestingly, when computing the period of strongest response in the viewing recordings, we saw high agreement between the different viewers. Further analysis revealed that different types

of facial features are useful for the prediction of different scores of emotional state. For example, we saw that simply counting the viewer's smiles and blinks (miscellaneous features) provided an inferior, yet significantly correlated, prediction of *Likability* and *Rewatch*. For commercial applications, these facial features can be obtained from the laptop's embedded camera (in a similar manner to [69]). Furthermore, we found that the dynamic aspects of facial expressions contribute more to the prediction of viewers affective state than to the prediction of media tags.

In a wider perspective, we recall that on April 2005, an 18 seconds video clip titled "Me at the zoo" became the first video clip uploaded to YouTube. In the decade since, the world had witnessed an unprecedented growth of online video resources – on YouTube alone there are over 1.2 billion video clips; adding other popular websites like Vimeo, Dailymotion and Facebook, and we reach an un-grasped amount of cuddling cats, soccer tricks and recorded DIY manuals. On April 2016 the CEO of Facebook, Mark Zuckerberg, stated that within 5 years Facebook will be almost entirely composed of videos; and with 1.8 billion active users that watch over 8 billion videos per day [3], that is a statement that should be taken seriously.

Such a staggering amount of videos poses many challenges to computer scientists and engineers. Since every user can upload videos as he desire, classifying and mapping them for accurate and quick retrieval, ease of use, search engines and recommendation systems becomes a challenging task. One solution is tagging the videos – assigning descriptive labels that aids indexing and arranging them. Another challenge is to understand the viewers' expected emotional response, to refine personal customization and to comprehend the effect of the videos on the users. These are the challenges we tackled in this work.



REVIEW OF EMOTION ELICIT DATABASES

The following pages includes a review of all major emotion eliciting databases. To be noted that the databases reviewed here might be partially to the ones released, as only the emotion eliciting clips are discussed. For example, both DEAP [52] and MAHNOB-HCI [91] contains major parts of subjects' physiological signals that are not mentioned in this review. Drawbacks are numbered with respect to Section 3.2

Reference	Description	Affect Descriptors	Drawbacks
LIRIS-ACCEDE [16]	9,800 excerpts extracted from 160 feature films and short films, shared under Creative Commons license. Duration of 8-12 seconds, rated online by over 2,000 crowdsourced participants.	Valence and Arousal on 5-point scale (derived from rating-by-comparison).	(1) Relatively short; (6) Familiarity wasn't examined; (7) Crowdsourced.
VideoEmotions [44]	1,101 YouTube (480) and Flickr (621) clips that were returned as search results for 8 different emotional adjectives (<i>e.g.</i> Anger, Joy and Disgust). Average duration of 107 seconds.	Search results labels: Anger, Anticipation, disgust, Fear, Joy, Sadness, Surprise, and Trust.	(1) 107 seconds in average; (2) No V-A ratings; (6) As above.
EMDB [19]	52 non-auditory 40-second excerpts from commercial films, rated by 113 subjects.	Valence, Arousal and Dominance on discrete 9-point scales, and free language text concerning felt emotion during and after watching.	(4) The lack of sound may be a confounding factor (because it's unusual) and alters the subjects' viewing experience; (6) Taken from well-known (partially Oscar winning) movies; (8) Most films are copyright protected.

DEAP [52]	120 1-minute excerpts (with maximum emotional content) from music video clips, rated by 14-16 subjects per clip.	Valence, Arousal and Dominance on discrete 9-point scales.	(1) Relatively long; (4) All clips are music videos. Although they differ in their emotional tags, they have many similarities (<i>e.g.</i> they are all accompanied by songs and dancing); (6) Many of them are clips of highly popular songs; (7) Rated by online subjects; (8) Some of the clips are not available online anymore due to copyright claims.
MAHNOB-HCI [91]	20 excerpts from Hollywood movies, YouTube clips and weather reports. Duration of 35-117 seconds ($\mu = 81.4$), rated by 27 participants.	Valence, Arousal, Dominance and Predictability on discrete 9-point scales, as well as basic emotion tagging.	(1) Relatively long; (3) Major sensor presence (EEG recordings, videotaping, audio recordings, physiological signals recording and eye tracking); (6) Most clips are from Hollywood famous movies; (8) Most films are copyright protected.
FilmStim [87]	70 excerpts from dozens of films selected by 50 film rental store managers. Duration of 1-7 minutes long, dubbed (or originally) in French, rated by a total of 364 subjects.	Arousal level and 16 emotional adjectives on a 7-point scales (<i>e.g.</i> Anxious, Tense, Nervous), Positive and Negative Affect Schedule (PANAS scale [105]).	(1) Over a minute long; (5) French Speaking; (6) Taken from well-known (partially Oscar winning) movies; (8) As above.

Untitled [102]	36 Full-length Hollywood movies (a total of 2040 scenes), rated by 3 subjects.	Most suitable emotion for each scene, out of 7 basic emotions.	(1) Assuming the average movie length is 1 hour and 30 minutes, the average scene length is about 95 seconds; (2) No V-A Rating; (6) As above; (8) As above.
Untitled [33]	78 excerpts extracted from over 250 films, 8-1192 seconds ($\mu = 151$), rated by at least 25 subjects per clip, with 494 subjects in total.	16 basic emotions on a 9-point scale.	(1) Average of 2:31; (2) As above; (6) As above; (8) As above.
Untitled [81]	12 excerpts extracted from 9 films. Duration of 3-6 minutes, dubbed (or originally) in French, rated by 60 subjects.	10 basic emotions on a 5-point scale, nine scales of semantic feeling (<i>e.g.</i> "Little – Large") and free labels.	(1) Relatively long; (5) French Speaking; (6) As above; (8) As above.

TABLE A.1: EMOTION ELICITING VIDEO CLIPS DATABASES.

BIBLIOGRAPHY

- [1] <http://msdn.microsoft.com/en-us/library/hh973074.aspx>.
- [2] <https://github.com/danielhadar/emotion-elicitedb>.
- [3] <https://techcrunch.com/2015/11/04/facebook-earnings-q3-2015/>.
- [4] <https://www.youtube.com/yt/policyandsafety/communityguidelines.html>.
- [5] <http://www.faceshift.com/>.
- [6] M. K. ABADI, S. M. KIA, R. SUBRAMANIAN, P. AVESANI, AND N. SEBE, *User-centric affective video tagging from meg and peripheral physiological responses*, in Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, IEEE, 2013, pp. 582–587.
- [7] S. AMAN AND S. SZPAKOWICZ, *Identifying expressions of emotion in text*, in International Conference on Text, Speech and Dialogue, Springer, 2007, pp. 196–205.
- [8] K. ANDERSON AND P. W. MCOWAN, *A real-time automated system for the recognition of human facial expressions*, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 36 (2006), pp. 96–105.
- [9] I. ARAPAKIS, I. KONSTAS, AND J. M. JOSE, *Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance*, in Proceedings of the 17th ACM international conference on Multimedia, ACM, 2009, pp. 461–470.
- [10] I. ARAPAKIS, Y. MOSHFEGHI, H. JOHO, R. REN, D. HANNAH, AND J. M. JOSE, *Enriching user profiling with affective features for the improvement of a multimodal recommender system*, in Proceedings of the ACM International Conference on Image and Video Retrieval, ACM, 2009, p. 29.
- [11] A. P. ATKINSON, M. L. TUNSTALL, AND W. H. DITTRICH, *Evidence for distinct contributions of form and motion information to the recognition of emotions from body gestures*, Cognition, 104 (2007), pp. 59–72.

- [12] X. BAO, S. FAN, A. VARSHAVSKY, K. LI, AND R. ROY CHOUDHURY, *Your reactions suggest you liked the movie: Automatic content rating via reaction sensing*, in Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, ACM, 2013, pp. 197–206.
- [13] S. A. BARGAL, E. BARSOUM, C. C. FERRER, AND C. ZHANG, *Emotion recognition in the wild from videos using images*, in Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 433–436.
- [14] M. S. BARTLETT, G. LITTLEWORT, M. FRANK, C. LAINSCSEK, I. FASEL, AND J. MOVELLAN, *Recognizing facial expression: machine learning and application to spontaneous behavior*, in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 2, IEEE, 2005, pp. 568–573.
- [15] ———, *Fully automatic facial action recognition in spontaneous behavior*, in 7th International Conference on Automatic Face and Gesture Recognition (FGR06), vol. 2, IEEE, 2006, pp. 223–230.
- [16] Y. BAVEYE, E. DELLANDREA, C. CHAMARET, AND L. CHEN, *Liris-accede: A video database for affective content analysis*, IEEE Transactions on Affective Computing, 6 (2015), pp. 43–55.
- [17] M. M. BRADLEY AND P. J. LANG, *Measuring emotion: the self-assessment manikin and the semantic differential*, Journal of behavior therapy and experimental psychiatry, 25 (1994), pp. 49–59.
- [18] D. H. BRAINARD, *The psychophysics toolbox*, Spatial vision, 10 (1997), pp. 433–436.
- [19] S. CARVALHO, J. LEITE, S. GALDO-ÁLVAREZ, AND Ó. F. GONÇALVES, *The emotional movie database (emdb): A self-report and psychophysiological study*, Applied psychophysiology and biofeedback, 37 (2012), pp. 279–294.
- [20] P. R. CHAKRABORTY, L. ZHANG, D. TJONDRONEGORO, AND V. CHANDRAN, *Using viewer’s facial expression and heart rate for sports video highlights detection*, in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 371–378.
- [21] C. H. CHAN AND G. J. JONES, *Affect-based indexing and retrieval of films*, in Proceedings of the 13th annual ACM international conference on Multimedia, ACM, 2005, pp. 427–430.
- [22] J. CHEN, B. HU, P. MOORE, X. ZHANG, AND X. MA, *Electroencephalogram-based emotion assessment system using ontology and data mining techniques*, Applied Soft Computing, 30 (2015), pp. 663–674.

- [23] T. DANISMAN AND A. ALPKOÇAK, *Feeler: Emotion classification of text using vector space model*, in AISB 2008 Convention Communication, Interaction and Social Intelligence, vol. 1, 2008, p. 53.
- [24] D. DE AMICI, C. KLERSY, F. RAMAJOLI, L. BRUSTIA, AND P. POLITI, *Impact of the hawthorne effect in a longitudinal clinical study: the case of anesthesia*, Controlled clinical trials, 21 (2000), pp. 103–114.
- [25] P. DELAVEAU, M. JABOURIAN, C. LEMOGNE, N. ALLAÏLI, W. CHOUCOA, N. GIRAULT, S. LEHERICY, J. LAREDO, AND P. FOSSATI, *Antidepressant short-term and long-term brain effects during self-referential processing in major depression*, Psychiatry Research: Neuroimaging, 247 (2016), pp. 17–24.
- [26] P. EKMAN, *An argument for basic emotions*, Cognition & emotion, 6 (1992), pp. 169–200.
- [27] P. EKMAN, W. V. FRIESEN, AND J. C. HAGER, *Facial action coding system (facs)*, A technique for the measurement of facial action. Consulting, Palo Alto, 22 (1978).
- [28] P. EKMAN AND E. L. ROSENBERG, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*, Oxford University Press, 1997.
- [29] T. W. FRAZIER, M. E. STRAUSS, AND S. R. STEINHAEUER, *Respiratory sinus arrhythmia as an index of emotional response in young adults*, Psychophysiology, 41 (2004), pp. 75–83.
- [30] A. D. GERSON, L. C. PARRA, AND P. SAJDA, *Cortically coupled computer vision for rapid image search*, IEEE Transactions on neural systems and rehabilitation engineering, 14 (2006), pp. 174–179.
- [31] S. B. GOKTURK, J.-Y. BOUGUET, C. TOMASI, AND B. GIROD, *Model-based face tracking for view-independent facial expression recognition*, in Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on, IEEE, 2002, pp. 287–293.
- [32] J. K. GOODMAN, C. E. CRYDER, AND A. CHEEMA, *Data collection in a flat world: The strengths and weaknesses of mechanical turk samples*, Journal of Behavioral Decision Making, 26 (2013), pp. 213–224.
- [33] J. J. GROSS AND R. W. LEVENSON, *Emotion elicitation using films*, Cognition & emotion, 9 (1995), pp. 87–108.
- [34] D. GRÜHN AND S. SCHEIBE, *Age-related differences in valence and arousal ratings of pictures from the international affective picture system (iaps): Do ratings become more extreme with age?*, Behavior Research Methods, 40 (2008), pp. 512–521.

- [35] J. GU, *Creating 3d images of objects by illuminating with infrared patterns*, Aug. 4 2009. US Patent 7,570,805.
- [36] H. GUNES, B. SCHULLER, M. PANTIC, AND R. COWIE, *Emotion representation, analysis and synthesis in continuous space: A survey*, in Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 827–834.
- [37] J. HAIDT AND D. KELTNER, *Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition*, Cognition & Emotion, 13 (1999), pp. 225–266.
- [38] J. HAN, X. JI, X. HU, L. GUO, AND T. LIU, *Arousal recognition using audio-visual features and fmri-based brain response*, IEEE Transactions on Affective Computing, 6 (2015), pp. 337–347.
- [39] S. H. HEMENOVER AND U. SCHIMMACK, *That’s disgusting!Ä, but very amusing: Mixed feelings of amusement and disgust*, Cognition and Emotion, 21 (2007), pp. 1102–1113.
- [40] H. C. HJORTSJÖ, *Man’s face and mimic language*, Student literature, 1969.
- [41] M. HUSSEIN AND T. ELSAYED, *Studying facial expressions as an implicit feedback in information retrieval systems*, (2008).
- [42] G. IRIE, T. SATOU, A. KOJIMA, T. YAMASAKI, AND K. AIZAWA, *Affective audio-visual words and latent topic driving model for realizing movie affective scene classification*, IEEE Transactions on Multimedia, 12 (2010), pp. 523–535.
- [43] X. JIANG, T. ZHANG, X. HU, L. LU, J. HAN, L. GUO, AND T. LIU, *Music/speech classification using high-level features derived from fmri brain imaging*, in Proceedings of the 20th ACM international conference on Multimedia, ACM, 2012, pp. 825–828.
- [44] Y.-G. JIANG, B. XU, AND X. XUE, *Predicting emotions in user-generated videos.*, in AAAI, 2014, pp. 73–79.
- [45] J. JIAO AND M. PANTIC, *Implicit image tagging via facial information*, in Proceedings of the 2nd international workshop on Social signal processing, ACM, 2010, pp. 59–64.
- [46] O. P. JOHN AND S. SRIVASTAVA, *The big five trait taxonomy: History, measurement, and theoretical perspectives*, Handbook of personality: Theory and research, 2 (1999), pp. 102–138.
- [47] H. JOHO, J. M. JOSE, R. VALENTI, AND N. SEBE, *Exploiting facial expressions for affective video summarisation*, in Proceedings of the ACM International Conference on Image and Video Retrieval, ACM, 2009, p. 31.

- [48] H. JOHO, J. STAIANO, N. SEBE, AND J. M. JOSE, *Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents*, Multimedia Tools and Applications, 51 (2011), pp. 505–523.
- [49] A. KAPOOR, P. SHENOY, AND D. TAN, *Combining brain computer interfaces with vision for object categorization*, in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.
- [50] M. KLEINER, D. BRAINARD, D. PELLI, A. INGLING, R. MURRAY, C. BROUSSARD, ET AL., *What's new in psychtoolbox-3*, Perception, 36 (2007), p. 1.
- [51] S. KOELSTRA, C. MÜHL, AND I. PATRAS, *Eeg analysis for implicit tagging of video data*, in 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, 2009, pp. 1–6.
- [52] S. KOELSTRA, C. MUHL, M. SOLEYMANI, J.-S. LEE, A. YAZDANI, T. EBRAHIMI, T. PUN, A. NIJHOLT, AND I. PATRAS, *Deap: A database for emotion analysis; using physiological signals*, IEEE Transactions on Affective Computing, 3 (2012), pp. 18–31.
- [53] S. KOELSTRA, M. PANTIC, AND I. PATRAS, *A dynamic texture-based approach to recognition of facial actions and their temporal models.*, IEEE transactions on pattern analysis and machine intelligence, 32 (2010), pp. 1940–54.
- [54] S. KOELSTRA AND I. PATRAS, *Fusion of facial expressions and eeg for implicit affective tagging*, Image and Vision Computing, 31 (2013), pp. 164–174.
- [55] P. J. LANG, *The emotion probe: Studies of motivation and attention.*, American psychologist, 50 (1995), p. 372.
- [56] P. J. LANG, M. M. BRADLEY, AND B. N. CUTHBERT, *International affective picture system (iaps): Affective ratings of pictures and instruction manual*, Technical report A-8, (2008).
- [57] J. T. LARSEN, A. P. MCGRAW, AND J. T. CACIOPPPO, *Can people feel happy and sad at the same time?*, Journal of personality and social psychology, 81 (2001), p. 684.
- [58] G. LEE, M. KWON, S. K. SRI, AND M. LEE, *Emotion recognition based on 3d fuzzy visual and eeg features in movie clips*, Neurocomputing, 144 (2014), pp. 560–568.
- [59] Y.-P. LIN, Y.-H. YANG, AND T.-P. JUNG, *Fusion of electroencephalographic dynamics and musical*, (2014).
- [60] C. L. LISETTI AND F. NASOZ, *Using noninvasive wearable computers to recognize human emotions from physiological signals*, EURASIP Journal on Advances in Signal Processing, 2004 (2004), pp. 1–16.

- [61] G. LITTLEWORT, M. S. BARTLETT, I. FASEL, J. SUSSKIND, AND J. MOVELLAN, *Dynamics of facial expression extracted automatically from video*, Image and Vision Computing, 24 (2006), pp. 615–625.
- [62] S. LUCEY, A. B. ASHRAF, AND J. F. COHN, *Investigating spontaneous facial action recognition through aam representations of the face*, INTECH Open Access Publisher, 2007.
- [63] C. D. MAESTAS AND J. POPE, *Subject know thyself? comparing self-reported and observed emotions and their influence on political attitudes*, Comparing Self-Reported and Observed Emotions and Their Influence on Political Attitudes (January 7, 2016), (2016).
- [64] I. B. MAUSS AND M. D. ROBINSON, *Measures of emotion: A review*, Cognition and emotion, 23 (2009), pp. 209–237.
- [65] S. M. MAVADATI AND M. H. MAHOOR, *Temporal facial expression modeling for automated action unit intensity measurement*, in Pattern Recognition (ICPR), 2014 22nd International Conference on, IEEE, 2014, pp. 4648–4653.
- [66] J. MCCAMBRIDGE, J. WITTON, AND D. R. ELBOURNE, *Systematic review of the hawthorne effect: new concepts are needed to study research participation effects*, Journal of clinical epidemiology, 67 (2014), pp. 267–277.
- [67] R. MCCARNEY, J. WARNER, S. ILIFFE, R. VAN HASELEN, M. GRIFFIN, AND P. FISHER, *The hawthorne effect: a randomised, controlled trial*, BMC medical research methodology, 7 (2007), p. 1.
- [68] D. MCDUFF, R. EL KALIOUBY, AND R. W. PICARD, *Crowdsourcing facial responses to online videos*, in Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, IEEE, 2015, pp. 512–518.
- [69] D. MCDUFF, R. EL KALIOUBY, T. SENECHAL, D. DEMIRDJIAN, AND R. PICARD, *Automatic measurement of ad preferences from facial responses gathered over the internet*, Image and Vision Computing, 32 (2014), pp. 630–640.
- [70] K. MCRAE, J. J. GROSS, J. WEBER, E. R. ROBERTSON, P. SOKOL-HESSNER, R. D. RAY, J. D. GABRIELI, AND K. N. OCHSNER, *The development of emotion regulation: an fmri study of cognitive reappraisal in children, adolescents and young adults*, Social cognitive and affective neuroscience, 7 (2012), pp. 11–22.
- [71] A. G. MONEY AND H. AGIUS, *Video summarisation: A conceptual framework and survey of the state of the art*, Journal of Visual Communication and Image Representation, 19 (2008), pp. 121–143.

- [72] R. NIESE, A. AL-HAMADI, A. PANNING, AND B. MICHAELIS, *Emotion recognition based on 2d-3d facial feature extraction from color image sequences*, Journal of Multimedia, 5 (2010), pp. 488–500.
- [73] D. PALOMBA, M. SARLO, A. ANGRILLI, A. MINI, AND L. STEGAGNO, *Cardiac responses associated with affective processing of unpleasant film stimuli*, International Journal of Psychophysiology, 36 (2000), pp. 45–57.
- [74] M. PANTIC AND M. BARTLETT, *Machine analysis of facial expressions*, in Face Recognition, K. D. Grgic and Mislav, eds., no. June, I-Tech Education and Publishing, 2007, ch. 20, pp. 377–419.
- [75] M. PANTIC AND I. PATRAS, *Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences*, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 36 (2006), pp. 433–449.
- [76] M. PANTIC AND L. J. ROTHKRANTZ, *Facial action recognition for facial expression analysis from static face images*, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 34 (2004), pp. 1449–1461.
- [77] M. PANTIC AND A. VINCIARELLI, *Implicit human-centered tagging [social sciences]*, IEEE Signal Processing Magazine, 26 (2009), pp. 173–180.
- [78] G. PAOLACCI, J. CHANDLER, AND P. G. IPEIROTIS, *Running experiments on amazon mechanical turk*, Judgment and Decision making, 5 (2010), pp. 411–419.
- [79] D. G. PELLI, *The videotoolbox software for visual psychophysics: Transforming numbers into movies*, Spatial vision, 10 (1997), pp. 437–442.
- [80] W.-T. PENG, C.-H. CHANG, W.-T. CHU, W.-J. HUANG, C.-N. CHOU, W.-Y. CHANG, AND Y.-P. HUNG, *A real-time user interest meter and its applications in home video summarizing*, in Multimedia and Expo (ICME), 2010 IEEE International Conference on, IEEE, 2010, pp. 849–854.
- [81] P. PHILIPPOT, *Inducing and assessing differentiated emotion-feeling states in the laboratory*, Cognition and emotion, 7 (1993), pp. 171–193.
- [82] G. SANDBACH, S. ZAFEIRIOU, M. PANTIC, AND L. YIN, *Static and dynamic 3d facial expression recognition: A comprehensive survey*, Image and Vision Computing, 30 (2012), pp. 683–697.
- [83] E. SARIYANIDI, H. GUNES, AND A. CAVALLARO, *Automatic analysis of facial affect: A survey of registration, representation, and recognition*, IEEE transactions on pattern analysis and machine intelligence, 37 (2015), pp. 1113–1133.

- [84] A. SAVRAN, B. SANKUR, AND M. T. BILGE, *Facial action unit detection: 3D versus 2D modality*, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, (2010), pp. 71–78.
- [85] A. SAVRAN, B. SANKUR, AND M. T. BILGE, *Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units*, Pattern recognition, 45 (2012), pp. 767–782.
- [86] A. SAVRAN, B. SANKUR, AND M. TAHA BILGE, *Regression-based intensity estimation of facial action units*, Image and Vision Computing, 30 (2012), pp. 774–784.
- [87] A. SCHAEFER, F. NILS, X. SANCHEZ, AND P. PHILIPPOT, *Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers*, Cognition and Emotion, 24 (2010), pp. 1153–1172.
- [88] J. D. J. SHOTTON, P. KOHLI, A. BLAKE, AND I.-E. GIVONI, *Image labeling with global parameters*, Oct. 14 2014.
US Patent 8,861,870.
- [89] F. SILVEIRA, B. ERIKSSON, A. SHETH, AND A. SHEPPARD, *Predicting audience responses to movie content from electro-dermal activity signals*, in Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, ACM, 2013, pp. 707–716.
- [90] M. SOLEYMANI, S. ASGHARI-ESFEDEN, Y. FU, AND M. PANTIC, *Analysis of eeg signals and facial expressions for continuous emotion detection*, IEEE Transactions on Affective Computing, 7 (2016), pp. 17–28.
- [91] M. SOLEYMANI, J. LICHTENAUER, T. PUN, AND M. PANTIC, *A multimodal database for affect recognition and implicit tagging*, IEEE Transactions on Affective Computing, 3 (2012), pp. 42–55.
- [92] M. SOLEYMANI AND M. PANTIC, *Human-centered implicit tagging: Overview and perspectives*, in 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2012, pp. 3304–3309.
- [93] ———, *Multimedia implicit tagging using eeg signals*, in 2013 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2013, pp. 1–6.
- [94] M. SOLEYMANI, M. PANTIC, AND T. PUN, *Multimodal emotion recognition in response to videos*, IEEE transactions on affective computing, 3 (2012), pp. 211–223.
- [95] R. M. A. TEIXEIRA, T. YAMASAKI, AND K. AIZAWA, *Determination of emotional content of video clips by low-level audiovisual features*, Multimedia Tools and Applications, 61 (2012), pp. 21–49.

- [96] Y.-L. TIAN, T. KANADE, AND J. F. COHN, *Recognizing action units for facial expression analysis*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23 (2001), pp. 97–115.
- [97] M. TKALCIC, A. ODIC, A. KOSIR, AND J. TASIC, *Affective labeling in a content-based recommender system for images*, IEEE transactions on multimedia, 15 (2013), pp. 391–400.
- [98] A. J. TOMARKEN, R. J. DAVIDSON, AND J. B. HENRIQUES, *Resting frontal brain asymmetry predicts affective responses to films.*, Journal of personality and social psychology, 59 (1990), p. 791.
- [99] T. TRON, A. PELED, A. GRINSPHOON, AND D. WEINSHALL, *Automated facial expressions analysis in schizophrenia: A continuous dynamic approach*, in International Symposium on Pervasive Computing Paradigms for Mental Health, Springer, 2015, pp. 72–81.
- [100] A. B. P. Y. O. TRON, TALIA AND PELED, ABRAHAM AND GRINSPHOON, ALEXANDER AND WEINSHALL, DAPHNA, TITLE=FACEAL EXPRESSIONS AND FLAT AFFECT IN SCHIZOPHRENIA, AUTOMATIC ANALYSIS FROM DEPTH CAMERA DATA.
- [101] T. VANDAL, D. MCDUFF, AND R. EL KALIOUBY, *Event detection: Ultra large-scale clustering of facial expressions*, in Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, vol. 1, IEEE, 2015, pp. 1–8.
- [102] H. L. WANG AND L.-F. CHEONG, *Affective understanding in film*, IEEE Transactions on circuits and systems for video technology, 16 (2006), pp. 689–704.
- [103] S. WANG, Z. LIU, Y. ZHU, M. HE, X. CHEN, AND Q. JI, *Implicit video emotion tagging from audiences,Â facial expression*, Multimedia Tools and Applications, 74 (2015), pp. 4679–4706.
- [104] S. WANG, Y. ZHU, G. WU, AND Q. JI, *Hybrid video emotional tagging using users,Â eeg and video content*, Multimedia tools and applications, 72 (2014), pp. 1257–1283.
- [105] D. WATSON, L. A. CLARK, AND A. TELLEGEN, *Development and validation of brief measures of positive and negative affect: the panas scales.*, Journal of personality and social psychology, 54 (1988), p. 1063.
- [106] R. WESTERMANN, G. STAHL, AND F. HESSE, *Relative effectiveness and validity of mood induction procedures: analysis*, European Journal of social psychology, 26 (1996), pp. 557–580.
- [107] M. XU, J. S. JIN, S. LUO, AND L. DUAN, *Hierarchical movie affective content analysis based on arousal and valence features*, in Proceedings of the 16th ACM international conference on Multimedia, ACM, 2008, pp. 677–680.

- [108] M. XU, J. WANG, X. HE, J. S. JIN, S. LUO, AND H. LU, *A three-level framework for affective content analysis and its case studies*, Multimedia Tools and Applications, 70 (2014), pp. 757–779.
- [109] S. YANG, M. KAFAI, L. AN, AND B. BHANU, *Zapping index: using smile to measure advertisement zapping likelihood*, IEEE Transactions on Affective Computing, 5 (2014), pp. 432–444.
- [110] A. YAZDANI, J.-S. LEE, AND T. EBRAHIMI, *Implicit emotional tagging of multimedia using eeg signals and brain computer interface*, in Proceedings of the first SIGMM workshop on Social media, ACM, 2009, pp. 81–88.
- [111] Z. ZENG, I. C. SOCIETY, M. PANTIC, AND S. MEMBER, *A Survey of Affect Recognition Methods : Audio , Visual , and Spontaneous Expressions*, 31 (2009), pp. 39–58.
- [112] M. ZENTNER, D. GRANDJEAN, AND K. R. SCHERER, *Emotions evoked by the sound of music: characterization, classification, and measurement.*, Emotion, 8 (2008), p. 494.
- [113] S. ZHAO, H. YAO, X. SUN, P. XU, X. LIU, AND R. JI, *Video indexing and recommendation based on affective analysis of viewers*, in Proceedings of the 19th ACM international conference on Multimedia, ACM, 2011, pp. 1473–1476.
- [114] Y. ZHU, S. WANG, AND Q. JI, *Emotion recognition from users' eeg signals with the help of stimulus videos*, in 2014 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2014, pp. 1–6.